Causation, Responsibility, and Typicality

Justin Sytsma[1]

Abstract: There is ample evidence that violations of injunctive norms impact ordinary causal attributions. This has struck some as deeply surprising, taking the ordinary concept of causation to be purely descriptive. Our explanation of the findings—the responsibility view—rejects this: we contend that the concept is in fact partly normative, being akin to concepts like responsibility and accountability. Based on this account, we predicted a very different pattern of results for causal attributions when an agent violates a statistical norm. And this pattern has been borne out by the data (Sytsma et al. 2012, Livengood et al. 2017, Sytsma ms-a). These predictions were based on the responsibility attributions that we would make. In this paper, I extend these previous findings, testing responsibility attributions. The results confirm the basis of our predictions, showing the same pattern of effects previously found for causal attributions for both injunctive norms and statistical norms. In fact, the results for responsibility attributions are not statistically significantly different from those previously found for causal attributions. I argue that this close correspondence lends further credence to the responsibility view over competing explanations of the impact of norms on causal attributions.

Injunctive norms have a notable impact on ordinary causal attributions (claims of the form "X caused Y").[2] To illustrate, consider the Pen Case from Knobe and Fraser (2008):

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly e-mailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith [illegible] ceptionist needs [illegible] left on her desk.

[1] Forthcoming in Review of Philosophy and Psychology: https://doi.org/10.1007/s13164-020-00498-2
I want to thank Joshua Knobe and Jonathan Kominsky for very helpful feedback on previous drafts of this paper.

[2] See, for example, Alicke (1992), Knobe and Fraser (2008), Hitchcock and Knobe (2009), Sytsma et al. (2012), Reuter et al. (2014), Kominsky et al. (2015), Livengood et al. (2017), Icard et al. (2017), Kominsky and Phillips (2019), Livengood and Sytsma (2020). Under "injunctive norms" I include both prescriptive norms (what should be done) and proscriptive norms (what should not be done), although "prescriptive norm" is often used to refer to both.

In this case, two agents perform symmetric actions, jointly bringing about a bad outcome. The only difference is that Professor Smith violated an injunctive norm (faculty members aren't supposed to take pens), while the administrative assistant did not (administrative assistants are allowed to take pens). When participants were asked to rate agreement with a causal attribution for each agent, however, ratings were very notably higher for Professor Smith.

How should we explain findings like this? In a series of papers my colleagues and I have argued that injunctive norms matter for ordinary causal attributions because the ordinary concept of causation at play in such attributions is a broadly moral concept, akin to the concepts of responsibility and accountability.[3] Our underlying motivation for this *responsibility view* is that human cognition is highly attuned to recognizing applicable injunctive norms (whether implicit or explicit), detecting and responding to violations, and navigating factors that exacerbate or mitigate those violations.[4] As Joshua Knobe (2010, 328) has put it, "we are moralizing creatures through and through." Taking moral considerations broadly, such that they include norms that fall well short of what we might want to term "moral," our view follows directly from this basic insight: if humans are (broadly) moralizing creatures through and through, then we should not find it surprising that a host of ordinary concepts are responsive to our broadly moral evaluations, including both causation and responsibility. Thus, while some have thought that causation is a purely descriptive matter, such that broadly moral evaluations must impact causal

[3] See Sytsma et al. (2012), Livengood et al. (2017), Sytsma et al. (2019), Livengood and Sytsma (2020), Sytsma and Livengood (ms), Sytsma (ms-a, ms-b, ms-c, ms-d). While discussions of the impact of injunctive norms on causal attributions often describe these in terms of moral judgments, this is arguably too strong. In keeping with Sytsma et al. (2012), I'll instead speak of broadly moral evaluations, where this is intended to correspond with the injunctive norms at issue. Finally, our thesis is first a claim about the ordinary use of causal attributions in English. Most of the work at issue in this paper have been conducted on English-speakers (but see Samland and Waldmann 2016, Grinfeld et al. 2020). How far our thesis extends to other languages is an interesting open question.

[4] That we are highly attuned to norms and their violation is hardly controversial (e.g., Sripada and Stich 2007), even as there is a great deal of disagreement about why we are so attuned and how this came about.

attributions *indirectly*, we deny this: we hold that evaluations of injunctive norms are *directly* relevant to our causal attributions.

In previous work (Sytsma et al. 2012, Livengood et al. 2017, Sytsma ms-a), we have shown what might seem like a surprising pattern of findings concerning the effects of another type of norm on causal attributions—statistical norms (i.e., whether an action or event is typical or atypical). A prominent indirect explanation of the impact of injunctive norms, the *counterfactual view* developed by Knobe and colleagues[5], predicts that statistical norms should also show the same type of effects on causal attributions. But we found a more complex pattern than this, with the results running counter to the predictions of the counterfactual view. In brief, we found that causal attributions for agents were insensitive to one type of statistical norm (what is typical or atypical for members of a relevant population to which the agent belongs); and we found that while causal attributions were sensitive to another type of statistical norm (what is typical or atypical for the agent herself) when the agent knows about the likely outcome of her action, the effect ran in the opposite direction to that predicted by the counterfactual view (causal ratings being higher when the agent acted typically than when she acted atypically).[6]

Our predictions about the impact of statistical norms on causal attributions were derived from thinking about responsibility, and as such we predict that similar effects should be seen for responsibility attributions. In this paper I test these predictions, repeating the studies from Livengood et al. (2017) and Sytsma (ms-a), but now asking participants about responsibility instead of causation. In line with our predictions, the complex pattern of effects for injunctive and statistical norms previously found for causal attributions is also found for responsibility attributions. In fact, the results for these two types of attributions are not statistically significantly

[5] See Hitchcock and Knobe (2009), Halpern and Hitchcock (2015), Kominsky et al. (2015), Phillips et al. (2015), Icard et al. (2017), Kominsky and Phillips (2019).
[6] Sytsma (ms-a, ms-b) and Sytsma and Livengood (ms) raise further issues for the counterfactual view.

distinguishable across these studies. This strengthens previous findings of a close correspondence between causal attributions and responsibility attributions (Sarin et al. 2017; Murray and Lombrozo 2017; Grinfeld et al. 2020; Sytsma and Livengood ms; Sytsma ms-a, ms-b, ms-d).[7] I argue that this close correspondence across both injunctive and statistical norms calls out for a common explanation and that the responsibility view offers the most plausible common explanation of these effects.

Here is how I will proceed. In Section 1, I discuss the previous findings for violations of statistical norms in the Pen Case and extend this to responsibility attributions. In Section 2, I extend the findings for the Lauren Alone Case from Livengood et al. (2017) to responsibility attributions. The implications of these results for explanations of ordinary causal attributions is discussed in Section 3.

1. The Pen Case

Knobe and Fraser (2008) found that causal ratings for the Pen Case were significantly higher for Professor Smith, who violates an injunctive norm, than for the administrative assistant, who does not violate a norm. Each of the main accounts in the literature is able to explain this effect, including the responsibility view and the counterfactual view.[8] The responsibility view offers a *direct explanation*, arguing that broadly moral evaluations are directly relevant to applying the

[7] In their first experiment, Grinfeld et al. (2020) used strength measures for responsibility and causation finding corresponding effects for a sample drawn from Amazon Mechanical Turk. By contrast, in their fourth experiment they found a predicted dissociation between strength measures of responsibility and causation based on epistemic perspective. It should be noted, however, that the study was conducted in Hebrew using university students, and that there were potentially important differences between the measures used, the responsibility question asking about the agent's responsibility for success or failure ("To what extent is the candidate responsible for his success/failure in the exam?") while the causal question asked about the agent's action with regard to simply the outcome ("To what extent did the study of the candidate cause the outcome in the exam?"). Further work is needed to confirm that the dissociation occurs in English and when using matching queries.

[8] In addition to these views, other prominent accounts in the literature include Alicke's *blame view* (e.g., Alicke 1992, Alicke et al. 2011) and Samland and Waldmann's *pragmatic view* (e.g., Samland and Waldmann 2016). I'll return to these views briefly in Section 3.

ordinary concept of causation at play in causal attributions. The counterfactual view, by contrast, offers an *indirect explanation*. This view holds that broadly moral evaluations influence the alternative possibilities we consider, which in turn play a role in our causal attributions. Specifically, norm violations are taken to make the counterfactual on which the norm-violation does not occur salient, such that people are more likely to consider it, and on that counterfactual in scenarios like the Pen Case the outcome would not occur, which leads participants to judge that the norm-violating agent is more causal. This account is not specific to injunctive norms, however, but holds that violations of any type of norm should impact causal attributions. Thus, the counterfactual view predicts that the same type of effect seen for violations of injunctive norms should also be found for violations of statistical norms. This has not been borne out by the data for scenarios involving agents, however.

While the cases focused on in this paper involve agents, the counterfactual and responsibility views are not specific to agents. There is some evidence that the impact of injunctive norms extends to what philosophers would normally consider non-agents (e.g., Hitchcock and Knobe 2009).[9] Further, there is some evidence suggesting that statistical norms impact causal attributions for non-agents (Kominsky et al. 2015, Icard et al. 2017, Morris et al. 2019). Kominsky et al.'s Experiment 4 and Icard et al.'s Experiment 2 did not test causal attributions, though, but "because" statements. And it is unclear whether such statements work analogously to the causal attributions—to "X caused Y" statements (Livengood and Machery 2007; Livengood et al. 2017, fn39). Morris et al. (2019) have recently reported similar findings testing causal attributions, however, although the attributions tested concerned events involving

[9] This effect is typically explained in terms of violations of a specific type of injunctive norm, what Hitchcock and Knobe term *norms of proper functioning*. We hold that such effects can be explained in terms of the application of the same normative concept at play in causal attributions concerning agents. In line with this, research indicates that people have a tendency to take an agentive perspective on nature as a whole, including what philosophers would think of as non-agents (e.g., Bloom 2007, Rose 2015, and Rose and Schaffer 2017).

an agent (Joe selecting a ball from a box in a game of chance), where the probability of success was varied. Further work is needed to determine whether these findings extend to direct attributions to non-agents and to patterns of behavior rather than probabilities of outcomes.[10]

1.1 Previous Findings

In Sytsma et al. (2012) we tested whether violations of statistical norms have the same effect on causal attributions for the Pen Case as violations of injunctive norms. We distinguished between two types of statistical norms: an agent could either act typically/atypically with regard to her own usual behavior (agent-level typicality/atypicality) or with regard to the usual behavior of a salient group that she belongs to (population-level typicality/atypicality). The impact of these two types of statistical norm were tested across five studies. We found that causal attributions were insensitive to the population-level statistical norm. And, while they were sensitive to the agent-level statistical norm, the impact went in the opposite direction to that predicted by the counterfactual view: agent-level *typicality*, not *atypicality*, corresponded with higher causal ratings for the norm-violating agent (Professor Smith). Further, these findings have recently been replicated in Study 3 in Sytsma (ms-a) using slight variations on eight key conditions from the original studies.[11]

[10] While more work is needed here, it is possible that statistical norms have different effects for causal attributions concerning non-agents than for agents. One possibility is that statistical norms might be taken to bear more directly on the injunctive norms typically taken to be applicable to non-agents. Here the focus has been on norms of proper functioning—what things should do in designed or otherwise regular systems—and violations of typicality might plausibly be understood as violations of such a norm (e.g., a component of a machine doing something other than its typical behavior would not just violate a statistical norm, but an injunctive norm insofar as the component does something it wasn't designed to do).

[11] I tested just the conditions from Sytsma et al. (2012) where the administrative assistant did not violate a norm, making a slight modification to the conditions where Professor Smith did not violate an injunctive norm: in the original study it was stated that the agents were *able* to take pens; in the replication it was instead stated that they were *allowed* to take pens to more clearly mark permissibility.

While the effects of statistical norms on causal attributions found in these studies run counter to the predictions of the counterfactual view, we predicted these effects on the basis of our alternative responsibility view. Reflecting on the responsibility attributions we would make, we predicted that violations of population-level statistical norms would have little impact on causal attributions for scenarios like the Pen Case. Our reasoning was that excuses like "everyone was doing it" aren't generally taken to notably mitigate responsibility for a bad outcome. In contrast, for agent-level statistical norms, we predicted that causal ratings would be higher when the action is typical for the agent *and* she knows that she is doing something she shouldn't and that a bad outcome might result. The idea is that when the agent could be expected to know that a bad outcome might result from her behavior, people would be more likely to see her as being responsible for the outcome when she typically acts in such a reckless way. This draws out that the responsibility view expects a number of factors beyond just injunctive norms to matter for causal attributions, including what the agent knew and what she desires, because factors like these matter for responsibility attributions (e.g., Cushman 2008, Gailey and Falk 2008, Lagnado and Channon 2008, Malle et al. 2014, Young and Saxe 2011). And there is a growing body of evidence that such factors also matter for causal attributions (Samland and Waldmann 2016, Kirfel and Lagnado 2017, Sytsma ms-c).

1.2 Study 1

Our predictions about the impact of statistical norms on causal attributions for the Pen Case were arrived at through reflecting on the responsibility attributions we would make. As such, assuming our responsibility judgments are typical, we would expect to find similar effects if we

were to directly test responsibility attributions. Testing this prediction for the Pen Case was the goal of my first study.

Each participant in Study 1 was given one of the eight variations on the Pen Case from Study 3 in Sytsma (ms-a). In the first four conditions, I varied whether Professor Smith violated an injunctive norm (permissible, impermissible) and whether she violated a population-level statistical norm (population typical, population atypical). In the last four, I again varied whether Professor Smith violated an injunctive norm, but this time also varied whether she violated an agent-level statistical norm (agent typical, agent atypical). In each of the eight conditions, the administrative assistant did not violate either norm: his action is always both permissible and typical, with this being specified at either at the population level (first four conditions) or the agent level (last four conditions). After reading the vignette, participants were asked to rate their agreement or disagreement with a responsibility attribution for each agent on a 7-point scale anchored at 1 with "strongly disagree," at 4 with "neutral," and at 7 with "strongly agree."[12] The attributions were presented in random order. Results were collected from 338 participants.[13]

[12] See Sytsma (ms-a) for vignettes. For the first four probes the attributions were "Professor Smith is responsible for the problem" and "The administrative assistant is responsible for the problem"; for the second four, the administrative assistant was named "John" and the attribution tested was "John is responsible for the problem."

[13] Participants for each study in this paper were recruited through advertising for a free personality test on Google. In addition to answering the questions reported below, participants were asked basic demographic questions and after the philosophical questions were given a 10-item Big Five personality inventory. In line with Sytsma (ms-a), participants were restricted to native English-speakers who were 16 years of age or older. Participants for Study 1 were 61.8% women (three non-binary), average age 34.4 years, ranging from 16 to 90. Given the higher percentage of women, the four ANOVAs described below were run with gender as a third between-participants factor. No significant gender effects were found.

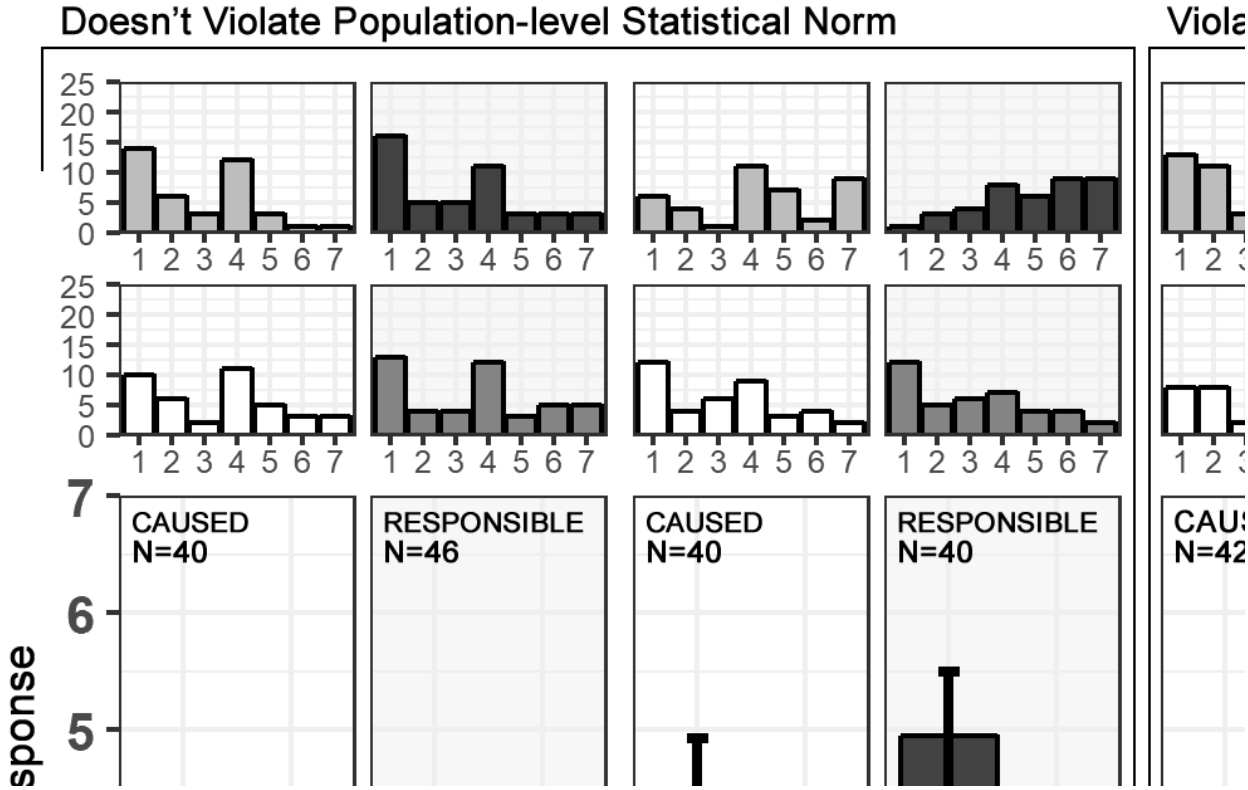

Figure 1: Results of first four conditions (population-level) for Sytsma (ms-a) Study 3 and Study 1; showing 95% confidence intervals, histograms shown above.

Results are shown in Figure 1 (population-level) and Figure 2 (agent-level) with those from Sytsma (ms-a) Study 3 shown for comparison. To look at the effect of term, I began by analyzing the results of Study 1 together with the results from Sytsma (ms-a), treating the population-level and agent-level probes separately and running ANOVAs with *term* (responsible, cause), *injunctive norm* (no violation, violation), and *statistical norm* (no violation, violation) as between-participant factors. Term was not significant for ratings of either Professor Smith or the administrative assistant for either set of probes.[14] For the population-level probes, statistical norm was not significant for ratings of either agent, while injunctive norm was significant for

[14] Professor Smith, Population-level: $F(1, 322)=1.34$, $p=0.25$, $\eta^2=0.003$. Administrative Assistant, Population-level: $F(1, 322)=0.18$, $p=0.67$, $\eta^2=0.001$. Professor Smith, Agent-level: $F(1, 327)=0.002$, $p=0.96$, $\eta^2=0.000$. Administrative Assistant, Agent-level: $F(1, 327)=2.82$, $p=0.094$, $\eta^2=0.008$.

both.[15] For the agent-level probes, both statistical norm and injunctive norm were significant for ratings of Professor Smith, while just statistical norm was significant for the administrative assistant.[16] There were no significant interaction effects in any of the ANOVAs.[17] And results were similar when analyzing just the results of Study 1.[18]

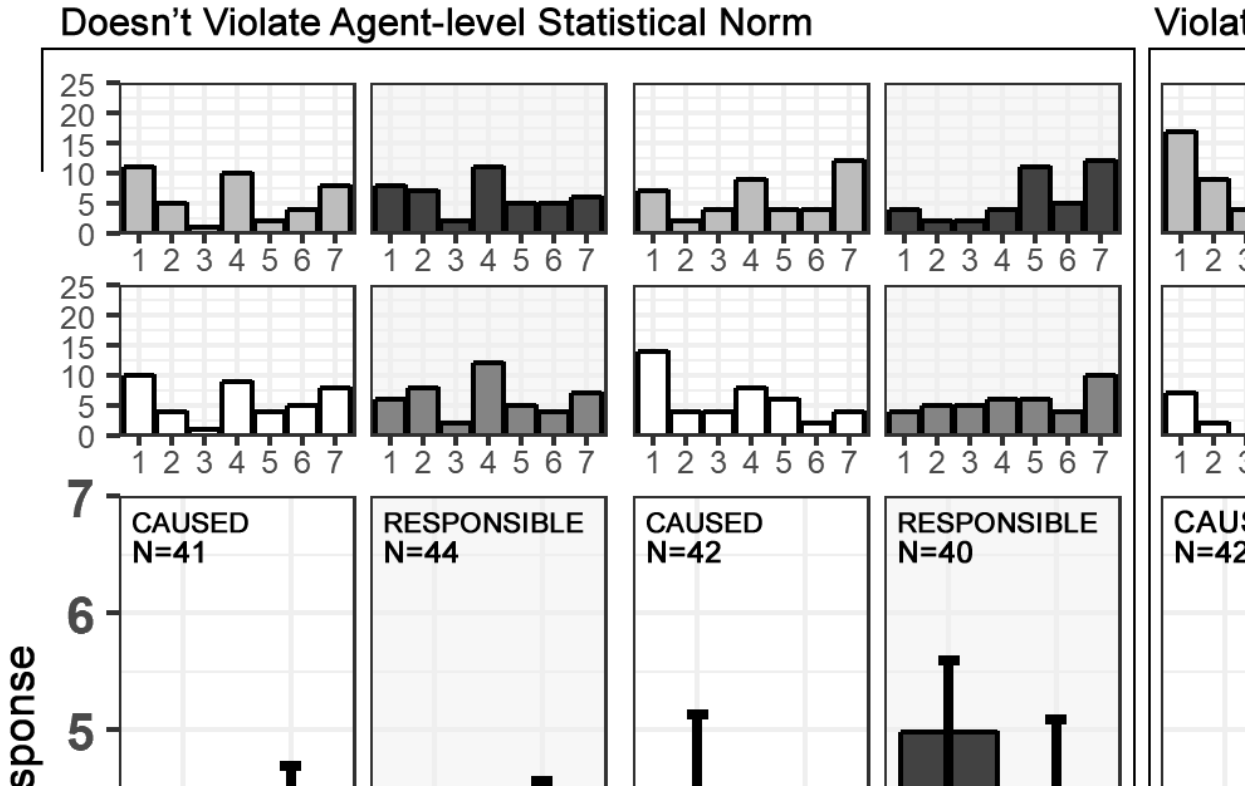

Figure 2: Results of last four conditions (agent-level) for Sytsma (ms-a) Study 3 and Study 1; showing 95% confidence intervals, histograms shown above.

[15] Professor Smith: statistical norm, $F(1, 322)=0.23$, $p=0.63$, $\eta^2=0.001$; injunctive norm, $F(1, 322)=68.3$, $p=3.7e^{-15}$, $\eta^2=0.17$. Administrative Assistant: statistical norm, $F(1, 322)=0.19$, $p=0.66$, $\eta^2=0.001$; injunctive norm, $F(1, 322)=4.07$, $p=0.045$, $\eta^2=0.012$.

[16] Professor Smith: statistical norm, $F(1, 327)=37.2$, $p=3.0e^{-9}$, $\eta^2=0.096$; injunctive norm, $F(1, 327)=19.6$, $p=1.3e^{-5}$, $\eta^2=0.051$. Administrative Assistant: statistical norm, $F(1, 327)=13.2$, $p=3.3e^{-4}$, $\eta^2=0.038$; injunctive norm, $F(1, 327)=1.64$, $p=0.20$, $\eta^2=0.005$.

[17] There was a borderline significant three-way interaction, however, for the Administrative Assistant in the agent-level conditions: $F(1, 327)=3.37$, $p=0.067$, $\eta^2=0.010$.

[18] Professor Smith, Population-level: injunctive norm, $F(1, 164)=34.1$, $p=2.8e\text{-}8$, $\eta^2=0.17$. Administrative Assistant, Population-level: no significant effects. Professor Smith, Agent-level: statistical norm, $F(1, 166)=30.5$, $p=1.2e\text{-}7$, $\eta^2=0.14$; injunctive norm, $F(1, 166)=13.8$, $p=2.7e\text{-}4$, $\eta^2=0.066$. Administrative Assistant, Agent-level: statistical norm, $F(1, 166)=4.34$, $p=0.039$, $\eta^2=0.025$; borderline significant interaction, $F(1, 166)=3.73$, $p=0.055$, $\eta^2=0.021$.

In line with the predictions of the responsibility view, we find that ratings for causal attributions and responsibility attributions for the Pen Case are remarkably similar. Further, violating the injunctive norm had a notable impact on ratings for Professor Smith. In Study 1, ratings were significantly higher for Professor Smith when she violated just the injunctive norm compared to when she violated neither norm for both the population- and agent-level conditions.[19] And ratings were significantly higher for Professor Smith in each pair of conditions when she violated both norms compared to when she violated just the statistical norm.[20] Comparable results were found for causal attributions in Sytsma (ms-a). Turning to the statistical norms, in line with our previous results for causal attributions, ratings were insensitive to the population-level statistical norm. And while responsibility attributions were sensitive to the agent-level statistical norm, ratings for Professor Smith were significantly *higher* for Professor Smith when she violated neither norm than when she violated just the agent-level statistical norm[21]; further, ratings were significantly *higher* for Professor Smith when she violated just the injunctive norm than when she violated both norms.[22] Again, comparable results were found for these comparisons for causal attributions in Sytsma (ms-a).

2. Lauren Alone Case

The standard scenarios that have been studied in the recent empirical literature on ordinary causal attributions involve two agents performing actions that are symmetric outside of one agent violating a norm. In Study 17 in Livengood et al. (2017), we diverged from this formula, testing

[19] Population-level: $t(83.986)=4.94$, $p=2.0e\text{-}6$, d=1.06, one-tailed; $W=1414.5$, $p=7.3e\text{-}6$. Agent-level: $t(81.84)=2.61$, $p=0.0054$, d=0.57, one-tailed; $W=1166$, $p=0.0047$.
[20] Population-level: $t(79.967)=3.49$, $p=4.0e^{-4}$, d=0.77, one-tailed; $W=1185.5$, $p=5.7e^{-4}$. Agent-level: $t(83.846)=3.16$, $p=0.0011$, d=0.67, one-tailed; $W=1254.5$, $p=0.0012$.
[21] $t(81.538)=4.23$, $p=3.0e^{-5}$, d=0.92, one-tailed; $W=443.5$, $p=2.8e^{-5}$.
[22] $t(83.404)=3.66$, $p=2.2e^{-4}$, d=0.79, one-tailed; $W=520.5$, $p=2.2e^{-4}$.

a scenario that featured just a single agent. In this section, I extend these findings to responsibility attributions.

2.1 Previous Findings

In the Lauren Alone Case a single agent is noted and her action is described as being sufficient to bring about a bad outcome. In these cases, Lauren works for a company that uses a mainframe computer. Unbeknownst to the company, the mainframe has recently become unstable, such that the system will crash if anyone logs into it. One day, Lauren logs into the mainframe and the system crashes. Participants were asked to rate whether they agreed or disagreed that Lauren caused the system to crash using the same 7-point scale as in Study 1. On a second page, participants were then told that the company learned about the problem with the mainframe and because of this implemented a policy prohibiting employees from logging in. Although Lauren knew about the policy, she once again logged into the mainframe and once again the system crashed. Participants were asked to rate the same causal attribution as on the first page.

We ran five variations on this scenario. In the first, no typicality information was provided.[23] In the remaining four, typicality information was provided on the first page, either describing Lauren's behavior as population typical, population atypical, agent typical, or agent atypical. We found that in each condition participants tended to deny that Lauren caused the system to crash when she didn't know about the problem and didn't violate a policy (first page) and to affirm that she caused the system to crash when she knew about the problem and violated a policy (second page). Overall, whether Lauren acted typically or atypically relative to either the population or her own behavior had no notable effect on participants' responses.

[23] The results for this condition have recently been replicated (Sytsma ms-c), including that the same effects were seen using a between-participants design with participants either receiving the probe from the first page or a slightly modified stand-alone version of the probe from the second page.

With regard to the Pen Case, in Sytsma et al. (2012) we predicted that agent typicality would have an impact on causal attributions. As noted above, however, the reasoning here involved Professor Smith knowing about the injunctive norm and the potential result of her behavior. But this is not the case for Lauren on the first page of the probes from Livengood et al. (2017): no injunctive norm is specified and there is little reason to suspect that Lauren knows about the problem with the mainframe.[24] As such, the responsibility view makes a different prediction about this case. In the absence of an injunctive norm and knowledge of the likely result of her action, we predict that causal attributions should also be insensitive to agent-level typicality. Again, our reasoning here was based on reflecting on the responsibility attributions we would make. As such, assuming our responsibility judgments are typical we would expect to find similar results when testing responsibility attributions for the Lauren Alone Case.

2.2 Study 2

Each participant in Study 2 was given one of the five variations on the Lauren Alone Case from Livengood et al. (2017). After reading the vignette on a given page, participants were asked to rate their agreement or disagreement with the responsibility attribution "Lauren is responsible for the system crashing" using the same 7-point scale as in Study 1.[25] Participants were not able to go back to first page after continuing to the second. Results were collected from 259 participants.[26]

[24] That participants don't infer that Lauren knows about the problem with the mainframe is confirmed by Studies 3 and 4 in Sytsma (ms-c).
[25] See Livengood et al. (2017) for vignettes.
[26] 69.1% women (six non-binary), average age 39.5 years, ranging from 16 to 86. The ANOVAs described below were run with gender as a second between-participants factor. No significant gender effects were found.

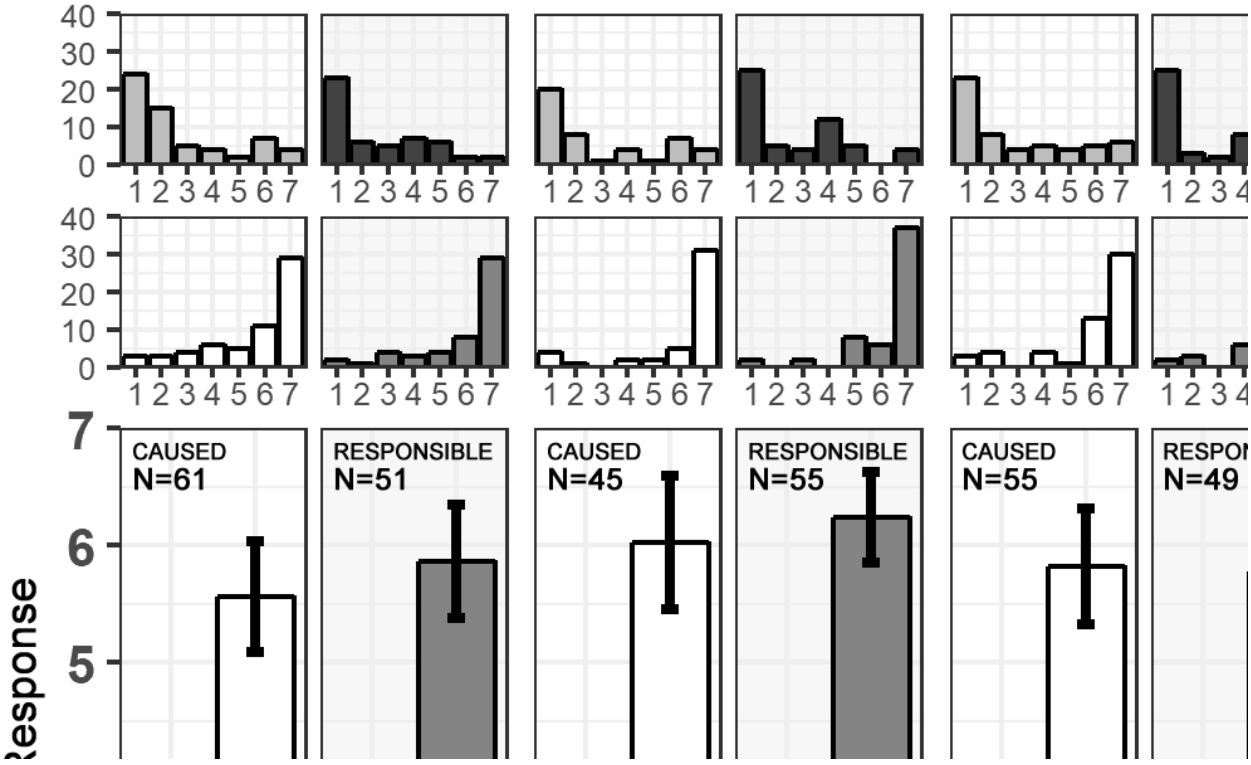

Figure 3: Results for Livengood et al. (2017) Study 17 and Study 2 with means followed by standard deviations below the bar graphs; showing 95% confidence intervals, histograms shown above. P1 show results for the probe where Lauren doesn't know about the issue and doesn't violate a policy, P2 for the probe where she knows about the problem and violates a policy.

Results are shown in Figure 3 with those from Livengood et al. (2017) Study 17 for comparison. To look at the effect of term, I began by analyzing the results of Study 2 together with the results from Livengood et al., running ANOVAs for responses on each page (no knowledge or policy, knowledge and policy) with *term* (responsible, cause) and *condition* (typicality not specified, agent-level typical, agent-level atypical, population-level typical, population-level atypical) as between-participant factors. No significant effects were seen for responses on either page.[27] Corresponding one-way ANOVAs for just the results of Study 2

[27] Page 1: Term, $F(1, 518)=1.25$, $p=0.26$, $\eta^2=0.002$; Condition, $F(4, 518)=1.10$, $p=0.36$, $\eta^2=0.008$. Interaction, $F(4, 518)=0.44$, $p=0.78$, $\eta^2=0.003$. Page 2: Term, $F(1, 518)=0.34$, $p=0.56$, $\eta^2=0.001$; Term*Condition, $F(4, 518)=1.79$, $p=0.13$, $\eta^2=0.013$; Term*Condition, $F(4, 518)=1.43$, $p=0.22$, $\eta^2=0.011$.

showed no significant effect of condition for responses on the first page (no knowledge or policy), but interestingly showed an effect for responses on the second page (knowledge and policy).[28]

Planned comparisons, however, indicate that the effect of condition for responses on the second page do not reflect the impact of either type of statistical norm; further, the overall pattern of effects matches that found by Livengood et al. for causal attributions. First, the mean response when Lauren doesn't know about the problem and doesn't violate a policy (first page) was significantly below the neutral point for each condition in Study 2, while the mean response when Lauren knows about the problem and violates a policy (second page) was significantly above the neutral point for each condition.[29] Second, comparing responses for the second and third conditions (agent-level typical versus agent-level atypical) for Study 2 showed no significant difference for either page, and similarly when comparing responses for the fourth and fifth conditions (population-level typical versus population-level atypical).[30] And the same held for causal attributions in Livengood et al.'s study.[31] These findings indicate that neither statistical norm has a notable effect on participants' responses on either the first page or the second page.

---

[28] Page 1: $F(1, 254)=0.79$, $p=0.53$, $\eta^2=0.012$. Page 2: $F(1, 254)=2.82$, $p=0.026$, $\eta^2=0.043$.

[29] Page 1, typicality not specified: $t(50)=5.29$, $p=1.4e^{-6}$, d=0.74, one-tailed; $V=131$, $p=6.8e^{-6}$. Page 2, typicality not specified: $t(50)=7.78$, $p=1.8e^{-10}$, d=1.09, one-tailed; $V=1079$, $p=9.4e^{-8}$. Page 1, agent-level typical: $t(49)=6.93$, $p=4.2e^{-9}$, d=0.98, one-tailed; $V=49$, $p=4.0e^{-7}$. Page 2, agent-level typical: $t(49)=4.14$, $p=6.9e^{-5}$, d=0.58, one-tailed; $V=843$, $p=3.8e^{-4}$. Page 1, agent-level atypical: $t(53)=7.09$, $p=1.6e^{-9}$, d=0.96, one-tailed; $V=71$, $p=8.3e^{-8}$. Page 2, agent-level atypical: $t(53)=5.27$, $p=1.3e^{-6}$, d=0.72, one-tailed; $V=898.5$, $p=2.5e^{-5}$. Page 1, population-level typical: $t(54)=5.10$, $p=2.3e^{-6}$, d=0.69, one-tailed; $V=141$, $p=1.6e^{-5}$. Page 2, population-level typical: $t(54)=11.6$, $p<2.2e^{-16}$, d=1.57, one-tailed; $V=1457$, $p=8.5e^{-10}$. Page 1, population-level atypical: $t(48)=3.30$, $p=9.2e^{-4}$, d=0.47, one-tailed; $V=225.5$, $p=0.0022$. Page 2, population-level atypical: $t(48)=6.93$, $p=4.8e^{-9}$, d=0.99, one-tailed; $V=861.5$, $p=4.4e^{-7}$.

[30] Page 1, agent-level typical vs. agent-level atypical: $t(101.95)=0.20$, $p=0.85$, d=0.038, two-tailed; $W=1391$, $p=0.78$. Page 2, agent-level typical vs. agent-level atypical: $t(101.42)=0.68$, $p=0.50$, d=0.13, two-tailed; $W=1205.5$, $p=0.33$. Page 1, population-level typical vs. population-level atypical: $t(92.697)=0.49$, $p=0.62$, d=0.098, two-tailed; $W=1342$, $p=0.97$. Page 2, population-level typical vs. population-level atypical: $t(91.502)=1.44$, $p=0.15$, d=0.29, two-tailed; $W=1511.5$, $p=0.22$.

[31] Page 1, agent-level typical vs. agent-level atypical: $t(105.98)=1.53$, $p=0.13$, d=0.29, two-tailed; $W=1209$, $p=0.12$. Page 2, agent-level typical vs. agent-level atypical: $t(104.4)=0.54$, $p=0.59$, d=0.10, two-tailed; $W=1585.5$, $p=0.35$. Page 1, population-level typical vs. population-level atypical: $t(93.331)=0.17$, $p=0.87$, d=0.034, two-tailed; $W=1205.6$, $p=0.81$. Page 2, population-level typical vs. population-level atypical: $t(93.261)=0.54$, $p=0.59$, d=0.11, two-tailed; $W=1384.5$, $p=0.25$.

Finally, I analyzed the combined results for the first three conditions, with the agent-level statistical norm as an ordered factor (atypical, unspecified, typical), and similarly for population-level typicality using the first, fourth, and fifth conditions. Again, no significant effects were seen for responses on either page.[32] And corresponding one-way ANOVAs for just the results of Study 2 showed no significant effect for either agent-level typicality or population-level typicality.[33]

3. General Discussion

The results from Sytsma et al. (2012), Livengood et al. (2017), and Sytsma (ms-a) suggest that while causal attributions to agents are sensitive to violations of injunctive norms, they are insensitive to violations of population-level statistical norms, and they are insensitive to violations of agent-level statistical norms when the agent is not prohibited from performing the action and does not know the likely outcome of her behavior. When the agent violates an injunctive norm and knows the likely result of her action, however, ordinary causal attributions are now sensitive to agent-level statistical norms, with causal ratings being *higher* when the agent acts *typically* than when she acts *atypically*.

These results for statistical norms run counter to the predictions of the counterfactual view but were predicted on the basis of our alternative responsibility view. Advocates of the counterfactual view could potentially respond, however, by modifying the view to drop the

[32] Agent-level, Page 1: Term, $F(1, 318)=1.35$, $p=0.25$, $\eta^2=0.004$; Norm, $F(2, 318)=0.75$, $p=0.47$, $\eta^2=0.005$; Term*Norm, $F(2, 318)=0.91$, $p=0.40$, $\eta^2=0.006$. Agent-level, Page 2: Term, $F(1, 318)=1.29$, $p=0.26$, $\eta^2=0.004$; Norm, $F(2, 318)=0.18$, $p=0.84$, $\eta^2=0.001$; Term*Norm, $F(2, 318)=2.13$, $p=0.12$, $\eta^2=0.013$. Population-level, Page 1: Term, $F(1, 310)=0.21$, $p=0.64$, $\eta^2=0.001$; Norm, $F(2, 310)=0.43$, $p=0.65$, $\eta^2=0.003$; Term*Norm, $F(2, 310)=0.031$, $p=0.97$, $\eta^2=0.000$. Population-level, Page 2: Term, $F(1, 310)=0.94$, $p=0.33$, $\eta^2=0.003$; Norm, $F(2, 310)=1.68$, $p=0.19$, $\eta^2=0.011$; Term*Norm, $F(2, 310)=0.28$, $p=0.76$, $\eta^2=0.002$.

[33] Agent-level, Page 1: $F(2, 152)=0.46$, $p=0.63$, $\eta^2=0.006$. Agent-level, Page 2: $F(2, 152)=1.77$, $p=0.17$, $\eta^2=0.023$. Population-level, Page 1: $F(2, 152)=0.24$, $p=0.79$, $\eta^2=0.003$. Population-level, Page 2: $F(2, 152)=1.17$, $p=0.31$, $\eta^2=0.015$.

general prediction about statistical norms. For instance, advocates might revise their account to hold that statistical norms do not directly affect the salience of alternative possibilities for agents, while still contending that the effect of injunctive norms on causal attributions works through counterfactual salience. And then the advocate of the counterfactual view might co-opt our explanation of the effect of agent-level statistical norms, taking such norms to operate on our broadly moral evaluations, which in turn affect counterfactual salience.

My primary goal here is not to further push the objection previously raised concerning statistical norms, however, but to raise a further objection for the counterfactual view based on the striking similarity between causal attributions and responsibility attributions. In the previous two sections, I tested the basis for our predictions in Sytsma et al. (2012) and Livengood et al. (2017), looking at responsibility attributions for the Pen Case and the Lauren Alone Case. As expected, the complex pattern of results previously found for causal attributions was also found for responsibility attributions. In fact, responsibility attributions were not statistically significantly distinguishable from causal attributions. This extends the growing body of evidence showing a striking similarity between causal attributions and responsibility attributions.

3.1 Common Explanation

Setting aside concerns based specifically on the impact of statistical norms on causal attributions, the close correspondence found between causal attributions and responsibility attributions calls for an explanation. There are two broad possibilities. First, it could be that there are separate mechanisms at play, one generating the pattern of effects for causal attributions and a different mechanism generating the pattern of effects for responsibility attributions. Second, it could be that there is a single mechanism, with the effects for both causal attributions and responsibility

attributions ultimately sharing a common cause. While it could be that separate mechanisms are at play but produce matching effects, such a coincidence is not overly plausible, and absent evidence to the contrary we should presume that there is a deeper connection. That is, we should favor a common explanation of these effects.

In fact, Knobe has argued for just such a conclusion with regard to the impact of broadly moral evaluations on causal attributions and other judgments, such as free action and intentionality (e.g., Knobe 2010, forthcoming; Phillips et al. 2015). For instance, in a recent paper he notes that research on people's judgments about these issues "has revealed something surprising," that "people's moral judgments appear to influence their judgments about all of these seemingly non-moral questions" (forthcoming, 1). Knobe then argues that while the effect of moral judgments in each of these areas might be due to unrelated processes, "given the obvious similarities between them, it is certainly tempting to seek a unified account" (6-7). I agree. And this reasoning holds especially strongly for the results for causation and responsibility, where we don't simply get similar effects, but statistically indistinguishable results for the exact same scenarios.

It might be argued that there is a difference between the findings for responsibility, on the one hand, and causation, free action, and intentionality on the other; specifically, that while the impact of broadly moral evaluations on the former is hardly surprising, the impact on the latter is quite surprising. That responsibility attributions are sensitive to injunctive norms is not surprising, nor following the reasoning laid out above, are the findings with regard to statistical norms: we would expect responsibility judgments to be sensitive to judgments concerning what

people should or shouldn't do.[34] In contrast, that injunctive norms have a corresponding impact on causal attributions has been taken to be surprising. For instance, after noting that it is not surprising that people's causal judgments would have an impact on their moral judgments, Knobe and Fraser (2008, 441, italics in original) argue that it is "more surprising… that the relationship can sometimes go *in the opposite direction*" with our moral judgments sometimes impacting our causal judgments.

Taking the results for responsibility to be unsurprising, and the results for causation, free action, and intentionality to be surprising, it might be thought that it is just the latter that call out for a common explanation. It is unclear why our surprise as theorists would warrant such a conclusion, however. If similarity of effects recommends seeking a common account, it would seem to do so even if we find some of those effects unsurprising. In fact, in such cases a strategy recommends itself—we often attempt to extend our explanation of the unsurprising result, which we would seem to have a better grip on, to explaining the surprising results. This is the type of strategy employed by the responsibility view: we hold that causal attributions are sensitive to injunctive norms for the same general reason that responsibility attributions are—i.e., that the ordinary concepts at play in these attributions have a normative component.

While we have not (yet) explicitly extended the responsibility view to judgments about free action or intentionality, it seems that a similar strategy could be applied. It might be that the concept of intentionality, for instance, is also best thought of as being a normative concept. In fact, some accounts in the literature on the effect of broadly moral considerations on judgments such as intentionality are similar to our view for causal attributions, such as Hindriks's (2008,

[34] While philosophers often distinguish between moral responsibility and causal responsibility (e.g., Talbert 2019), the impact of injunctive norms on responsibility attributions in the scenarios tested would seem to indicate that participants were calling on a concept more like the former.

2014) normative reasons explanation.[35] While the same basic type of strategy applies here, it is likely that the relationship between these concepts is more complex. For example, Hindriks (2014, 56) plausibly suggests that intentionality judgments play a role in our responsibility attributions, hypothesizing that intentionality judgments facilitate attributions of responsibility. Further, there is evidence that intentionality judgments emerge quickly and impact the subsequent formation of blame judgments (Monroe and Malle 2017).

Insofar as the broad type of account we offer can be extended to related effects in the literature, offering a common explanation does not favor the counterfactual view over the responsibility view. Knobe writes:

> Why might people's judgments about these apparently non-moral questions be influenced by moral considerations? A variety of different hypotheses immediately suggest themselves. Perhaps the effect can be explained in terms of people's emotional reactions, or perhaps it can be explained in terms of motivated cognition, or in terms of conversational pragmatics. (forthcoming, 1)

Extending the responsibility view, our proposed explanation is more straightforward than these and more straightforward than Knobe's preferred explanation in terms of thinking about alternative possibilities: the basic explanation is that appearances are sometimes deceiving and that these questions are not in fact generally non-moral, even if many philosophers have assumed that they are purely descriptive. And such an explanation would seem applicable not just to causation, but also to free action and intentionality.

[35] Hindriks explains the effect for intentionality judgments in terms of a misalignment between what should motivate the agent and the agent's actual reasons for acting, and he argues that the same holds for responsibility attributions. He then takes his explanation to provide reason to reject the call for a common explanation of the effects for intentionality and other folk-psychological concepts, on the one hand, and causal attributions on the other. The reason Hindriks offers is that "causing is a more objective notion deciding and acting intentionally," such that "it would be implausible if the attitudes of the acting agent bore systematically on what she caused" (2014, 67). Interestingly, Sauer (2014) arrives at the opposite conclusion for basically the same reason. Accepting the call for a common explanation, he holds that we should reject accounts that would not apply to causal attributions, assuming for instance that "whether or not an agent had a certain type of causal impact on the unfolding events does not depend on the described agents' beliefs" (491). There is evidence, however, that whether or not the agent knows that the outcome will occur if she acts impacts causal attributions (e.g., Sytsma mc-c).

3.2 The Role of Broadly Moral Evaluations

Focusing on the impact of broadly moral evaluations on causal attributions and responsibility attributions, I contend that (a) we should favor a common explanation for both, and that (b) for such an explanation we should favor one that extends the unsurprising findings to the surprising findings. This is the type of explanation that the responsibility view offers: we hold that our broadly moral evaluations are relevant to both our causal attributions and our responsibility attributions. The idea is that people tend to form broadly moral evaluations of a situation, and that such evaluations are then called on in making a range of attributions, including both causal attributions and responsibility attributions.

Our explanation contrasts with the counterfactual view, which holds that the concept at play in causal attributions is purely descriptive and explains the effect of broadly moral evaluations indirectly through the effect they have on the alternative possibilities that we consider. As discussed in Sytsma (ms-a), the responsibility view does not deny that counterfactual thinking plays some role in people's judgments; what it does deny, however, is that the impact of broadly moral evaluations on causal attributions (and responsibility attributions) works primarily through counterfactual salience.

One tempting possibility is that the alternative possibilities we consider play a role in our broadly moral evaluations and that such evaluations then play a role in both causal attributions and responsibility attributions. In fact, in recent work on modal cognition, Phillips and colleagues (Phillips and Cushman 2017, Phillips et al. 2019) have presented evidence suggesting this type of account. For instance, Phillips and Cushman (2017, 4650) summarize:

> this evidence suggests that there may be a nondeliberative and early-emerging form of modal cognition that incorporates information about both descriptive and prescriptive norms. According to this hypothesis, the default representation would support not only

> judgments of what is "possible," but also diverse modal judgments concerning what "ought," "should," "might," "may," or "could" be done.

Further, they hold that "these representations of alternative possibilities support many important cognitive functions, such as predicting others' future actions, assigning responsibility for past events, and making moral judgments" (4649). While this work is treated as being in line with the counterfactual view, it plausibly suggests the reverse ordering between broadly moral evaluations and the alternative possibilities we consider—how we represent alternative possibilities shaping our broadly moral evaluations rather than the other way around. Although the responsibility view does not make a specific suggestion concerning the role of the alternative possibilities we consider on our broadly moral evaluations, adopting thus suggestion, the simple models in Figure 4 show the key difference between the responsibility view (model **[1]**) and the counterfactual view (model **[2]**).

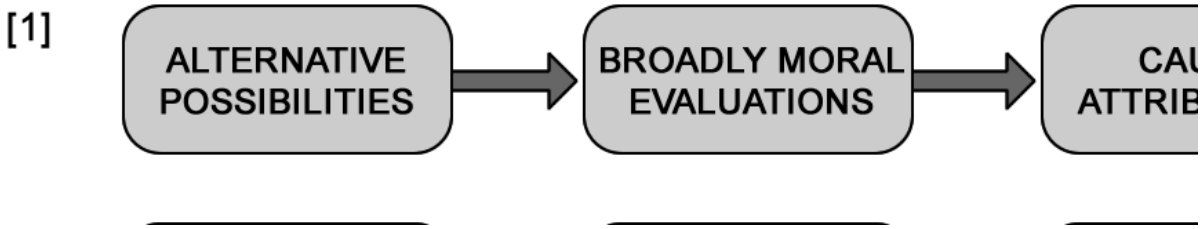

Figure 4: Two simple models for the impact of broadly moral evaluations on causal attributions, **[1]** compatible with the responsibility view and **[2]** the counterfactual view.

We can give corresponding models for responsibility attributions, as shown in Figure 5. For responsibility attributions, the model corresponding with the counterfactual view seems decidedly implausible. Model **[2*]** holds that broadly moral evaluations do not primarily impact responsibility attributions directly, but instead work through the more general mechanism of counterfactual salience. This type of model is plausible insofar as one takes the concept at issue

to be purely descriptive, as many philosophers have done for causal attributions, since the indirect mechanism preserves the descriptive nature of the concept, shielding off the normative considerations. When applied to a normative concept like that at play in responsibility attributions, however, this seems to get things the wrong way around: the motivation for an indirect mechanism does not apply in this case, since broadly moral evaluations seem directly relevant to applying the concept. Put another way, the impact of broadly normative considerations on responsibility attributions is hardly surprising, and this suggests in favor of the direct explanation offered by model **[1*]**.

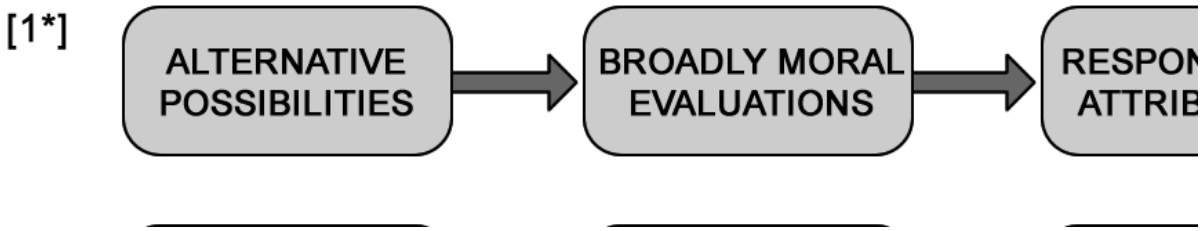

Figure 5: Two simple models for the impact of broadly moral evaluations on responsibility attributions.

Accepting (a)—accepting that we should explain the effect of broadly moral evaluations on causal attributions and responsibility attributions in the same way—we should favor corresponding models, either **[1]** and **[1*]** or **[2]** and **[2*]**. In line with (b), I contend that we should favor the pair that captures the unsurprising finding. That is, we should favor the models compatible with the responsibility view. The reason is simply that the direct explanation of the effect for responsibility attributions captures why this result is unsurprising. Given this, rejecting the direct explanation for responsibility attributions seems more revisionary than accepting the direct explanation for causal attributions.

3.3 Alternative Model

There is a potential conflict between the work of Phillips and colleagues noted above and how the counterfactual view is typically expressed. The counterfactual view has generally been laid out in terms of the impact of moral *judgments* on counterfactual salience and subsequently causal attributions, as seen above in the quotes from Knobe and Fraser (2009) and Knobe (forthcoming). Similarly, Phillips et al. (2015, 40) write: "At the core of our account is a claim about the impact of moral judgments on intuitions about the relevance of alternative possibilities." In the more recent work from Phillips and colleagues (Phillips and Cushman 2017, Phillips et al. 2019), however, they argue that how we represent alternative possibilities impacts our moral judgments, including our responsibility judgments, as well as our causal judgments. Following this alternative ordering, the counterfactual view might be updated to focus on the impact of *normative information* (information pertinent to our broadly moral evaluations) rather than our broadly moral evaluations themselves. It might then be argued that normative information impacts the alternative possibilities we consider, which in turn impacts our causal attributions, responsibility attributions, and broadly moral evaluations, as shown in Figure 6.[36]

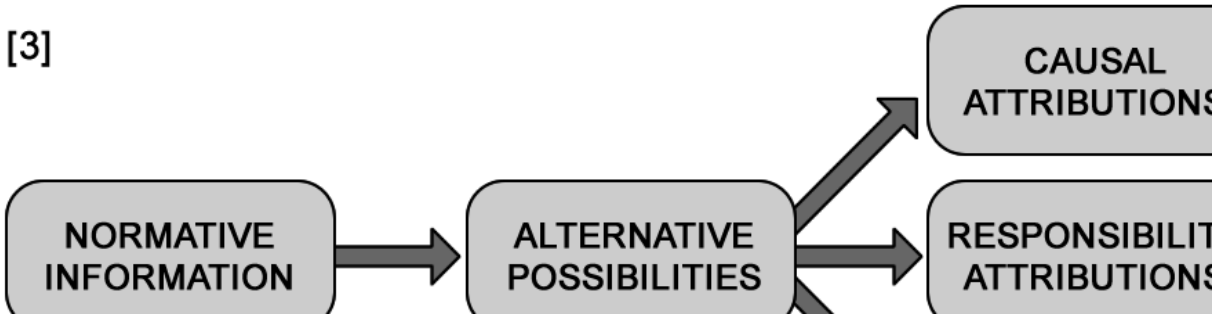

Figure 6: Simple alternative model for the counterfactual view.

[36] Alternatively, some of the attributions might be thought to run in sequence, for instance with the effect of alternative possibilities on responsibility attributions running through causal attributions. Further, it is worth noting that one would expect other factors to play a role in some attributions. For instance, Cushman (2008) presents evidence that blame judgments are far more sensitive to consequences and how they came about than judgments about the wrongness of an agent's action. We would expect these findings for blame attributions to extend to causal attributions and responsibility attributions, and plausibly the counterfactual view would make the same prediction.

As noted above, the key difference between the counterfactual view and the responsibility view is that the former takes the concept at play in causal attributions to be purely descriptive, while the latter denies this, holding that the concept also has a normative element. On model **[3]**, however, it is unclear how to apply this distinction. Insofar as the impact of normative information on responsibility attributions and broadly moral evaluations works through the alternative possibilities we represent, applying the same reasoning as for causal attributions would render both types of judgments purely descriptive as well. But then the distinction seems to lose all purchase: if the distinction is to mean anything, then surely it makes no sense to treat responsibility attributions and broadly moral evaluations as purely descriptive. Rather, what this model would seem to suggest is that information pertinent to our broadly moral evaluations has a deep influence on our thinking, including on our causal attributions. And this seems better described as drawing out the way in which each of these attributions has a normative element, rather than treating them all as purely descriptive. Noting the normative dimension of these concepts, however, is in line with our responsibility view.

Perhaps it might be argued that for causal attributions the influence of normative information reflects error—the normative information inappropriately influencing the application of a purely descriptive concept—while for responsibility attributions and broadly moral evaluations it reflects the appropriate application of this information. But this would seem to beg the question. Given the striking similarity between causal attributions and responsibility attributions, treating the sensitivity of one to normative information as an error but not the other seems unwarranted. If the (supposed) error is so widespread and systematic as to render the use of the descriptive concept statistically indistinguishable from the use of the normative concept, I contend that this isn't best treated as an error at all.

Arguably the most compelling evidence in favor of the counterfactual view—the manipulation studies reported by Phillips et al. (2015) and Kominsky and Phillips (2019)—can be interpreted in line with model **[3]**. In Phillips et al.'s second experiment they gave participants the neutral version of the Pen Case (the version in which neither agent violates a norm), then varied whether participants were asked to briefly describe what Professor Smith could have done differently or were asked to briefly describe the scenario. The idea is that writing about what Professor Smith could have done differently will increase the salience of the alternative actions that he could have taken, inducing the same effect previously found for violation of the injunctive norm. And the same method is employed by Kominsky and Phillips's third experiment, using a modified version of the Pen Case. In each case, they found the predicted effect: causal ratings were higher when participants wrote about what Professor Smith could have done differently. That this effect is found when directly manipulating counterfactual salience in the absence of information about injunctive norm violations would seem to offer strong support for the link from alternative possibilities to causal attributions in models **[2]** and **[3]**, although they interpret the studies in line with model **[2]**.

It is important to note, however, that in these studies Phillips and colleagues did not test participants' broadly moral evaluations. While Phillips et al. state that the claim that moral judgments impact intuitions concerning alternative possibilities is at the core of their account, as noted above, they do not actually test moral judgments, but instead assume that they follow directly from the materials presented. Thus, it is an open question whether their results reflect the impact of people's broadly moral evaluations on the relevance of alternative possibilities (as in model **[2]**) or whether the relevance of alternative possibilities instead impacts their broadly moral evaluations (as in models **[1]** and **[3]**). This open question was tested in a recent

replication of Phillips et al.'s study (Sytsma ms-d) by including a moral evaluation ("It was wrong for Professor Smith to take a pen."). The same manipulation effect was found for the moral evaluation, with wrongness judgments being significantly higher when participants wrote about what Professor Smith could have done differently. Further a Greedy Equivalence Search produced model **[1]**, with the effect of the manipulation on causal attributions running through the moral evaluation. In addition to supporting model **[1]**, these results push against model **[3]**, suggesting that the influence of alternative possibilities on our causal attributions runs through our broadly moral evaluations. In other words, the results are in line with model **[3*]** in Figure 7, which is compatible with the responsibility view.[37]

Figure 7: Alternative to model **[3]** that is compatible with the responsibility view.

3.4 Bias and Pragmatics

While I have focused on the counterfactual and responsibility views in this paper, these are not the only explanations of the impact of broadly moral considerations on causal attributions in the literature. Other notable accounts include Alicke's *bias view* and Samland and Waldmann's *pragmatic view*. Both views hold that the impact of injunctive norms on causal attributions

[37] As noted above, the responsibility view does not make a specific suggestion concerning the role of the alternative possibilities we consider in arriving at broadly moral evaluations, but is open to the possibility that they play a role. Similarly, we consider the exact relationship between causal attributions and responsibility attributions to be an open question. For instance, it could be that these run in sequence. In fact, in Sytsma (ms-d) participants were also asked to rate a responsibility attribution. Including this question in the search produced the same model, but with the effect of the wrongness judgments on causal attributions running through responsibility attributions.

reflects that participants are committing an error. The bias view holds that participants' judgments tend to be biased by their desire to blame (or praise) the agent, while the pragmatic view holds that participants tend to interpret the causal attributions in the experiments at issue as instead asking about responsibility or accountability.

Both of these views would predict some correspondence between causal attributions and responsibility attributions, taking the bias or pragmatic factors to lead participants to rate causal attributions in a similar way to how they would rate responsibility attributions. For the cases at issue, however, we don't just see *some* correspondence between causal attributions and responsibility attributions, but that they are statistically indistinguishable. Explaining this *close* correspondence on either the bias or pragmatic views would require positing an extremely strong and general error. In effect, to explain the present results these views would need to contend that basically all we are seeing here is error. But we would want very strong evidence indeed to accept such a conclusion, and currently there is no such evidence that I am aware of favoring one or the other of these views over the more charitable responsibility view. In fact, the present evidence otherwise supports the responsibility view over the bias view (Sytsma ms-c) and the pragmatic view (Sytsma et al. 2019).

4. Conclusion

Knobe (2010, 320) notes that "people's ordinary application of a variety of different concepts can be influenced by moral considerations," including their causal attributions. To this list we should add, rather unsurprisingly, people's responsibility attributions. The evidence presented in this paper indicates that not only are responsibility attributions impacted by broadly moral considerations, but they are strikingly similar to causal attributions for the scenarios tested. This

includes that they show the same pattern of effects for statistical norms previously found for causal attributions—effects that run counter to the counterfactual view developed by Knobe and colleagues. Thus, insofar as we have reason to posit a common explanation for the effects of broadly moral considerations on the application of concepts like causation and intentionality, as Knobe contends, we have reason to extend this to responsibility as well.

In attempting to explain the effect of broadly moral considerations on our judgments, Knobe notes two basic types of approaches we might take:

> One approach would be to suggest that moral considerations actually figure in the competencies people use to understand the world. The other would be to adopt what I will call an *alternative explanation*. That is, one could suggest that moral considerations play no role at all in the relevant competencies, but that certain additional factors are somehow 'biasing' or 'distorting' people's cognitive processes and thereby allowing their intuitions to be affected by moral judgments. (320)

Our responsibility view offers an explanation of the first type, holding that broadly moral evaluations are directly relevant to the application of the concepts at play in causal attributions and responsibility attributions. And while we have focused on these two types of judgments, our account might be extended to others that are sensitive to broadly moral considerations. The underlying motivation for the responsibility view is in line with the reasons Knobe offers for preferring the first type of approach—that we are (broadly) moralizing creatures through and through. Our account follows directly from this, holding that if this is the case then it should not be at all surprising that a host of ordinary concepts in part track our broadly moral concerns.

In contrast, the common explanation that Knobe has developed might be thought to fall short of this initial motivation. He has argued that a range of judgments, including causal judgments, are impacted by the alternative possibilities that we consider, with those possibilities in turn being impacted by normative judgments. Thus, Knobe distances the relevant judgments from broadly moral evaluations in two ways. First, he posits that it is norms generally, not just

injunctive norms, that matter. Second, he posits that norms primarily influence the relevant judgments indirectly. We have seen that there are problems for the first posit with regard to causal attributions. Focusing on injunctive norms, however, the second posit is plausible insofar as we take the impact of broadly moral considerations on the relevant judgments to be surprising (as has been common with regard to causal attributions), but it is less plausible when broadly moral considerations seem directly relevant to our judgments (as they do for responsibility attributions). As such, accepting that a common explanation is to be preferred, I have argued that we should favor extending the natural explanation of the unsurprising results to the surprising ones. We should set aside our theoreticians' surprise that ordinary causal attributions would be used in a way that marks our broadly moral evaluations and apply the most direct explanation—that causal attributions are akin to responsibility attributions in having a normative component.

References

Alicke, M. (1992). "Culpable causation." *Journal of Personality and Social Psychology*, 63: 368–378.

Alicke, M., D. Rose, and D. Bloom (2011). "Causation, Norm Violation, and Culpable Control." *Journal of Philosophy*, 108: 670–696.

Bloom, P. (2007). "Religion is Natural." *Developmental Science*, 10: 147–151.

Cushman, F. (2008). "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment." *Cognition*, 108: 353–380.

Gailey, J. and R. Falk (2008). "Attribution of Responsibility as a Multidimensional Concept." *Sociological Spectrum*, 28: 659–680.

Grinfeld, G., D. Lagnado, T. Gerstenberg, J. Woodward, and M. Usher (2020). "Causal Responsibility and Robust Causation." Frontiers in Psychology, 11: 1069.

Halpern, J. and C. Hitchcock (2015). "Graded Causation and Defaults." *British Journal for the Philosophy of Science*, 66: 413–457.

Hindriks, F. (2008). "Intentional Action and the Praise-Blame Asymmetry." *Philosophical Quarterly*, 58(233): 640–641.

Hindriks, F. (2014). "Normativity in Action: How to Explain the Knobe Effect and its Relatives." *Mind & Language*, 29(1): 51–72.

Hitchcock, C. and J. Knobe (2009). "Cause and Norm." *The Journal of Philosophy*, 106: 587–612.

Icard, T., J. Kominsky, and J. Knobe (2017). "Normality and Actual Causal Strength." *Cognition*, 161: 80–93.

Kirfel, L. and D. Lagnado (2017). "'Oops, I did it Again': The Impact of Frequency on Causal Judgements." *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Knobe, J. (2010). "Person as Scientist, Person as Moralist." Behavioral and Brain Sciences, 33: 315–365.

Knobe, J. (forthcoming). "Morality and Possibility." In J. Doris and M. Vargas (eds.), *The Oxford Handbook of Moral Psychology*. Oxford: Oxford University Press.

Knobe, J. and B. Fraser (2008). "Causal judgments and moral judgment: Two experiments." In W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447, Cambridge: MIT Press.

Kominsky, J., J. Phillips, T. Gerstenberg, D. Lagnado, and J. Knobe (2015). "Causal superseding." *Cognition*, 137: 196–209.

Kominsky, J. and J. Phillips (2019). "Immoral Professors and Malfunctioning Tools: Counterfactual Relevance Accounts Explain the Effect of Norm Violations on Causal Selection." *Cognitive Science*, 43(11): e12792.

Lagnado, D. and S. Channon (2008). "Judgments of Cause and Blame: The Effects of Intentionality and Foreseeability." *Cognition*, 108: 754–770.

Livengood, J. and E. Machery (2007). "The Folk Probably Don't Think What You Think They Think: Experiments on Causation by Absence." *Midwest Studies in Philosophy*, 31: 107–127.

Livengood, J., J. Sytsma, and D. Rose (2017). "Following the FAD: Folk attributions and theories of actual causation." *Review of Philosophy and Psychology*, 8(2): 274–294.

Livengood, J. and J. Sytsma (2020). "Actual causation and compositionality." *Philosophy of Science*, 87(1): 43-69.

Malle, B., S. Guglielmo, and A. Monroe (2014). "A Theory of Blame." *Psychological Inquiry*, 25: 147–186.

Monroe, A. and B. Malle (2017). "Two Paths to Blame: Intentionality Directs Moral Information Processing Along Two Distinct Tracks." *Journal of Experimental Psychology: General*, 146(1): 123–133.

Morris, A., J. Phillips, T. Gerstenberg, and F. Cushman (2019). "Quantitative Causal Selection Patterns in Token Causation." *PLOS One*, 14(8): e0219704.

Murray, D. and T. Lombrozo (2017). "Effects of Manipulation on Attributions of Causation, Free Will, and Moral Responsibility." *Cognitive Science*, 41: 447-481.

Phillips, J., J. Luguri, and J. Knobe (2015). "Unifying Morality's Influence on Non-moral Judgments: The Relevance of Alternative Possibilities." *Cognition*, 145: 30–42.

Phillips, J. and F. Cushman (2017). "Morality constrains the default representation of what is possible." *PNAS*, 114(18): 4649-4654.

Phillips, J., A. Morris, and F. Cushman (2019). "How We Know What Not To Think." *Trends in Cognitive Sciences*, 23(12): 1026–1040.

Reuter, K., L. Kirfel, R. van Riel, and L. Barlassina (2014). "The good, the bad, and the timely: How temporal order and moral judgment influence causal selection." *Frontiers in Psychology*, 5: 1336.

Rose, D. (2015). "Persistence Though Function Preservation." *Synthese*, 192: 97–146.

Rose, D. and Schaffer, J. (2017). “Folk mereology is teleological.” Noûs, 51(2): 238–270.

Sarin, A., D. Lagnado, and P. Burgess (2017). “The Intention-Outcome Asymmetry Effect: How Incongruent Intentions and Outcomes Influence Judgments of Responsibility and Causality.” *Experimental Psychology*, 64(2): 124-141.

Samland, J. and M. R. Waldmann (2016). “How prescriptive norms influence causal inferences.” *Cognition*, 156: 164–176.

Sauer, H. (2014). “It’s the Knobe Effect, Stupid! How (and How Not) to Explain the Side-Effect Effect.” *Review of Philosophy and Psychology*, 5:x485–503.

Sripada, C. and S. Stich (2007). “A Framework for the Psychology of Norms.” In P. Carruthers, S. Laurence, and S. Stich (eds.), *The Innate Mind Vol 2: Culture and Cognition*, pp. 280–301, New York: Oxford University Press.

Sytsma, J. (ms-a). “Structure and Norms: Investigating the Pattern of Effects for Causal Attributions.” http://philsci-archive.pitt.edu/16626/

Sytsma, J. (ms-b). “The Effects of Single versus Joint Evaluations on Causal Attributions.” http://philsci-archive.pitt.edu/16678/

Sytsma, J. (ms-c). “The Character of Causation: Investigating the Impact of Character, Knowledge, and Desire on Causal Attributions.” http://philsci-archive.pitt.edu/16739/

Sytsma, J. (ms-d). “Resituating the Influence of Relevant Alternative on Attributions.” http://philsci-archive.pitt.edu/16957/

Sytsma, J. and J. Livengood (ms). “Causal Attributions and the Trolley Problem.” http://philsci-archive.pitt.edu/16200/

Sytsma, J., J. Livengood, and D. Rose (2012). “Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions.” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43: 814–820.

Sytsma, J., R. Bluhm, P. Willemsen, and K. Reuter (2019). “Causal Attributions and Corpus Analysis.” In E. Fischer and M. Curtis (eds.), *Methodological Advances in Experimental Philosophy*, London: Bloomsbury Press.

Talbert, M. (2019). “Moral Responsibility.” In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition).

Young, L. and R. Saxe (2011). “When Ignorance is No Excuse: Different Roles for Intent Across Moral Domains.” *Cognition*, 120: 202–214.